Don’t @ Me: Experimentally Reducing Partisan Incivility on Twitter

Kevin Munger*

Prepared for 2016 Toronto Political Behaviour Workshop; currently unfinished.

Abstract

I conduct an experiment which examines the differential impact of several forms of moral suasion on partisans engaged in uncivil arguments. Partisans often respond in vitriolic and unproductive ways to tweets from politicians they disagree with, and this often engenders hateful responses from supporters of that politician. This phenomenon was especially common during the contentious 2016 US Presidential Election. Using Twitter accounts that I controlled, I sanctioned people engaged in this kind of behavior. By varying the language of the sanctioning, and the partisan identities of my accounts, I test hypotheses motivated by Moral Foundations Theory about how best to discourage toxic behavior online.

1 Introduction

There is a general impression that political discourse today is less civil than it was twenty years ago, and that changing norms related to the style of cable news is partially responsible (Berry and Sobieraj, 2013; Mutz, 2015). Scholars and the public alike are

*Department of Politics, New York University, 19 West 4th Street, 2nd floor, New York, NY, USA. email: km2713@nyu.edu.
concerned that this incivility degrades the quality of political discussion and can cause people to withdraw from engaging in politics.

Even more recently, social media has become a major platform for news distribution, elite communication and mass political discussion. Although there is less scholarly work examining the impact of online civility on the political process (but see Theocharis et al. (2015)), there is a huge amount of concern about online harassment. The harassment of women and minorities, for example, can have a chilling effect on their online participation (Henson, Reynolds, and Fisher, 2013; Hinduja and Patchin, 2007; Mantilla, 2013). This phenomenon has sufficiently permeated the Western cultural consciousness that no less august an institution than the long-running television series *South Park* has dedicated an entire season to exploring the de-mobilizing effects of online harassment.

*South Park* is one of many observers to tie this trend to the US presidential candidacy of Donald Trump. Trump has frequently decried “political correctness,” and many of his supporters cite his willingness to say exactly what he thinks without regard for violating speech norms as a major part of his appeal. Some of his statements are widely regarded as offensive to women, racial and religious minorities, and the disabled. Publicizing these views may have emboldened others who share them to express similarly offensive content online, leading to a concerted effort by members of the “alt right” to harass people with whom they disagree (Gross, 2016; Posner, 2016).

A similar trend has been observed with people on the opposite end of the political spectrum, sometimes called the “intolerant left.” Indeed, a major topic of discussion immediately before Trump’s campaign began dominating the headlines was the campus culture created by students’ attempts to dictate the range of tolerable discourse in the name of promoting social justice (Lukianoff and Haidt, 2015). Although their aims, tactics and moral frameworks differ from those on the alt right, online social justice activists have also engaged in coordinated efforts to harass people with whom they disagree (Ronson, 2016).

In this paper, I attempt to categorize and measure these behaviors in the context of the 2016 US Presidential election. I test different methods for encouraging civility online and evaluate their effectiveness. Using a method developed in an earlier paper (Munger, 2016), I used Twitter accounts that I controlled to sanction users engaged in uncivil discussions on Twitter. I sample users by searching for tweets that mention either @realDonaldTrump or @HillaryClinton but which are directed at another, non-elite user. Using an algorithm developed to identify aggression in comments on a Wikipedia editors’ discussion forum (Wulczyn and Thain, 2016), I selected the tweets most likely
to be uncivil. I then manually inspected the interaction to ensure that it was a true instance of a non-elite\(^1\) being uncivil to another non-elite of an opposing political persuasion. I then randomly assigned the subject to a treatment arm–subject to balance constraints–and used “bots” to send them a message.

By manipulating the partisan identity of my “bots”,\(^2\) I test the differential effects of sanctioning on Republicans and Democrats, as well as the degree of overlap between Democrats/Hillary supporters and Republicans/Trump supporters. By varying the language I tweet at subjects, I test hypotheses about the relative effectiveness of two kinds of moral suasion and include a “placebo check” by including a treatment arm with a message with no sanctioning.\(^3\)

I intend to analyze two forms of outcomes: the way the subjects respond to the tweet my bot sent at them, and the way that subjects’ behavior changes in the months after the treatment. I conducted the experiment in early October, so I only have results of the first kind. (And I don’t have them just yet–my apologies, but the experiment took longer to conduct than I had anticipated. I will have preliminary results by next week’s conference)

2 The Promise and Perils of Social Media

Perceptions of the impact of social media (and the internet more generally) on democratic politics have changed dramatically in the brief period of social media’s existence. Initial optimism suggested that citizens would be better able to communicate with both their governments and with each other, unconstrained by geography and the power imbalances of the physical world (Papacharissi, 2002). Although conversations could get heated and impolite, the overall effect was to revitalize the public sphere of debate (Papacharissi, 2004). The campaign manager for Howard Dean, one of the first politicians in the US to fully embrace the power of the Internet for politics, said that “the Internet is the most democratizing innovation we’ve ever seen, more so even than the printing press” (Trippi (2004), quoted in Hindman (2008)).

\(^1\)I define as an “elite” anyone who was “Verified” on Twitter–they have a blue check mark next to their name which means that Twitter has verified that they are who they say there, a status which Twitter only bestows on users they consider public figures–or anyone who identified themselves as a journalist or political operative in their profile.

\(^2\)These are not “bots” in the sense that they behave autonomously; I did all of the tweeting manually. I refer to them as bots throughout the paper for lack of a better term.

\(^3\)All hypotheses were pre-registered at EGAP.org prior to any research activities.
Indeed, a wide variety of politicians began using social media to communicate with their constituents (Gulati and Williams, 2010). Individual politicians are better able to reach voters directly, rather than through the mediating institution of party control (Karlsen and Skogerbø, 2013). Although the process does not always work perfectly, there is evidence that politicians respond to the citizens who engage with them on social media, discussing topics that citizens bring to their attention (Barberá et al., 2014). Additionally, citizens do seem to learn about party platforms directly from communication by politicians on Twitter (Munger et al., 2016).

On the non-elite side, the use of the internet to discuss non-political topics has enabled some cross-cutting ideological mass discussion (Wojcieszak and Mutz, 2009). This phenomenon first began with blogs. By 2006, 8 million US citizens claimed to share their thoughts through online blogs, and fully 57 million US citizens claimed to read them (Hindman (2008), p104). Hindman describes the prevailing mood at that time, when media commentators were lauding the development of blogs as a brave new world for deliberative democracy: “The central claim about blogs is that they amplify the political voice of ordinary citizens.” However, as he argues persuasively in The Myth of Digital Democracy, the infrastructure of the internet tends to lead to an even more skewed distribution of readership than does traditional media: “It may be easy to speak in cyberspace, but it remains difficult to be heard. (p142)”

When the competition to be heard is intense, competitors often resort to using outrageousness to garner attention. For example, when cable enabled new entrants to the television marketplace, these upstart media organizations were willing to blend news and entertainment in a way that traditional network broadcasters had resisted. In the words of Bill O’Reilly, host of the famously confrontational television program The O’Reilly Factor: “The best [cable news] host is the guy or gal who can get the most listeners extremely annoyed over and over and over again” (O’Reilly (2003), cited in Mutz (2015)). Norms of journalistic integrity established in the early 20th century rapidly eroded, resulting in less civil media and citizens who trusted and liked that media less (Berry and Sobieraj, 2013; Ladd, 2011).

A similar trend took place in citizen online engagement, but more rapidly and to a greater extreme. Early forums tended to be anonymous, and early internet users flocked to sites like 4chan and somethingawful to discuss whatever was on their mind. However, a subset of these people found that this anonymity empowered them to say uncivil and outrageous things, and that they could easily upset other users. This behavior soon spread over the internet, as “trolls” mocked memorial pages on Facebook and posted
vivid images of gore and hardcore pornography so that other users might suffer serious emotional turmoil (Phillips, 2015).

This kind of behavior is only possible through Computer Mediated Communication (CMC). In the physical world, biological feedback mechanisms make it emotionally difficult to look a stranger in the eye and say something uncivil (Frijda, 1988), but these mechanisms are lacking in CMC, as are physical proximity and identifiability. CMC makes it difficult to enforce social norms, and while this does tend to encourage more communication and creativity, it also allows even a small number of ill-intentioned actors to impose significant emotional costs on other users (Bordia, 1997; Kiesler, Siegel, and McGuire, 1984; Walther, 1996).

The competition for attention and the difficulty of punishment in anonymous contexts meant a race-to-the-bottom in terms of online speech norms, and the Internet is widely regarded as rife with offensive and even harassing speech designed to mock sincere expression–trolling culture is dominant online (Buckels, Trapnell, and Paulhus, 2014; Milner, 2013). The extent to which trolling culture obtains, though, depends on the specific technical affordances of different online platforms. The most important feature, in this respect, is the extent to which platforms allow their users to be anonymous. Studies have consistently found that the more anonymous platforms experience more harassment (Hosseinmardi et al., 2014; Omernick and Sood, 2013).

Facebook, for example, has invested heavily in linking their users’ accounts with their real identities. Twitter, on the other hand, allows all manner of parody, comedy and anonymous accounts. Twitter has consistently defined itself as in favor of free speech, and while this has made it the preferred platform for revolutionaries in both Western countries and authoritarian regimes around the world (Barberá et al., 2015; Earl et al., 2013), it has also become notorious for failing to curtail harassment. In the candid words of Twitter’s CEO Dick Costelo in an internal memo in 2015, “We suck at dealing with abuse and trolls on the platform and we’ve sucked at it for years.”

3 Social Media and Affect Polarization

The development of social media as both a platform for political communication and a locus for incivility took place at the same time as a sharp growth in animosity between Democratic and Republican partisans. Scholars have described this trend as “affect polarization”–partisans dislike each other (Iyengar, Sood, and Lelkes, 2012) and tend
to trust co-partisans and distrust out-partisans more (Iyengar and Westwood, 2015). This phenomenon has even extended to the marriage market, as preferences for a partner with similar partisan characteristics is stronger than ever (Huber and Malhotra, 2013).

Although the uptick in partisan polarization began well before the mass adoption of social media, there exists a plausible connection between the two. Some scholars claim that social media use exposes people to a wider range of views and thus decreases issue polarization (Barberá, 2014), but others argue that social media inflames partisan emotions and increases affect polarization (Settle, Forthcoming). The large-scale, contemporaneous development of social media and affect polarization makes causal claims difficult to establish; an exception is Lelkes, Sood, and Iyengar (2015), who use the quasi-random rollout of broadband internet as an instrument for the use of social media and finds that it significantly increased affect polarization.

In some ways, incivility is entailed by increasing affect polarization. I follow Mutz (2015), who draws a connection between civility and following the norms of politeness in a given society: “Following the rules of civility/politeness is...a means of demonstrating mutual respect. (p7)” If mutual respect between partisans is decreasing, it should be no surprise that civility in their conversations is decreasing as well.

Regardless of causality, it is clear that uncivil political arguments take place on social media. Sometimes the incivility is directed at politicians themselves, and while we might expect that having a thick skin is necessary to survive in that business, Theocharis et al. (2015) show that this can decrease politician engagement with their constituents on Twitter. Perhaps more importantly for the mass public, this behavior means that citizens who wish to engage with politicians or each other in response to a politicians’ tweet are necessarily exposed to uncivil messages.

### 4 Experimentally Reducing Political Incivility

Although Twitter has made efforts to reduce the incidence of incivility and harassment, it remains a large problem. Building on previous work to experimentally reduce racist harassment on Twitter (Munger, 2016), I conduct an experiment to sanction users who are sending uncivil messages to out-partisans and measure the change in their behavior.

The first step in performing this experiment was finding conversations that were uncivil, between out-partisans, and about politics. As in my previous experiment, where I searched for racist harassment by scraping tweets containing the slur “n****r,”
I first attempted to use a keyword search. I could not figure out a term that would reliably find the interactions I was looking for.

Instead, I used streamR to scrape the streaming Twitter API for tweets mentioning either “@realDonaldTrump” or “@HillaryClinton”—the Twitter accounts of the two major party candidates in the 2016 US Presidential election. I then dropped any tweets that were not directed at another user who was not either Trump or Clinton. Sending an uncivil message to Twitter accounts managed by teams of campaign workers is not exactly morally laudable—it is perhaps akin to muttering obscenities at a campaign ad played on an airport television—but it is less important from a deliberative point of view.

This way, I found a sample of tweets from non-elites that were concerned with the “issues” most likely to inspire political incivility in October 2016: Trump and Clinton. In order to filter through the hundreds of thousands of tweets every hour that fit these criteria, I used a machine learning classifier developed by Wulczyn and Thain (2016) to detect aggression. Wulczyn and Thain trained and evaluated a neural network on millions of comments on Wikipedia “talk pages” (the behind-the-scenes part of Wikipedia where editors discuss potential changes) in a format that is reasonably similar in structure and length to Tweets.

I used the model to assign an “aggression score” to each tweet I had scraped, then manually evaluated the top 10% most aggressive tweets per batch. From these prospective subjects, I selected the ones who were directing incivil language at a member of the opposite political persuasion. Many of the potential subjects I found this way were tweeting at elites—either people verified on Twitter, journalists or campaign operatives—and I excluded them. I also found many people agreeing (though often in uncivil ways) with an in-partisan about how terrible the out-party is, and excluded them as well. When performing a manual inspection of the potential subject’s profile, I excluded users who appeared to be minors or who were not tweeting in English. I also checked to ensure that the subjects’ profile was at least two months old; Twitter does ban some user accounts for harassment or other violations of their Terms of Service, so a very new account is likely to have been started by someone who had previously been banned. A new user is also likely to have too short a tweeting history for me establish a reasonable baseline for their past behavior.

For a visual overview of this selection process, see Figure 1. In this way, I found

---

4This process is time-consuming, and there were a finite number of tweets satisfying my criteria being tweeted at a given time, so I iterated this scrape-validate-treat procedure several times.
uncivil tweets from non-elites to non-elites with whom they disagreed politically. For an example, see Figure 2. @realDonaldTrump tweeted something, then Parker tweeted “you already lost” at Trump. Ty then responded to Parker (but because of how Twitter works, Ty’s tweet also “mentions” @realDonaldTrump) with an incivil comment. Ty is the subject I include in the experiment, and because he was being incivil to someone criticizing Trump, I coded Ty as a Trump supporter.

Based on findings in my previous experiment, and on the theoretical expectation that anonymity is an essential part of what enables incivility online, I also recorded each subject’s Anonymity Score during the subject discovery process. The Anonymity Score ranged from 0 (least anonymous, full name and picture) to 2 (most anonymous, no identifying information). Ty, from Figure 2, was coded as a 1—he chose to display what could plausibly be his full name. He also provided some personal information in his “bio” field, to the left of where he claims to be an “All around nice guy!”, which I censor for privacy reasons.

My aim was to convince subjects that they were being sanctioned by real person, so I made my bots look as real as possible. After I tweeted at a subject, they received a “notification” from Twitter. Non-elites are unlikely to get more than a few notifications per day, so they almost certainly saw the message I sent them. It is uncommon to be tweeted at by a stranger, but not extremely so, and especially not among a subject pool who are tweeting uncivil things at out-partisans. As a result, they were likely to click on my bots’ profile; if they did, they would see something very like Figure 3.

Neil, in panel (a), was a bot who appeared to be pro-Clinton. I created four bots; the other three were pro-Democrats, pro-Trump, and pro-Republicans (see Todd, in panel (b)). To change these identities, I changed the large banner in the middle of the profile, the small logo in the bottom right of the bots’ profile pictures, and the “bio” field below their username (eg “Hillary 2016!”; “Republicans 2016!”). The four bots were otherwise identical. All of the bots appeared to be white men, keeping race/gender aspect of the treatment constant. I used identical cartoon avatars to avoid anything about the users’ appearance priming the subjects; it is not uncommon for Twitter users to have cartoon avatars, so this was unlikely to raise suspicions.

I took other steps in order to maximize verisimilitude. Most importantly, I ensured that all of the bots had a reasonably high number of followers. In Munger (2016), I varied the number of followers that sanctioning bots had, and found that bots with

—I censor the usernames of the subjects to preserve their anonymity. In principle, the exact text of a tweet should be enough to find a user, but the phrases used in this exchange are quite common.
Figure 1: Sample Selection Process

Yes

StreamR finds a tweet with “@realDonaldTrump” or “@HillaryClinton”

Is the tweet an “@”-reply to someone besides Trump or Clinton?

Calculate aggression score; is tweet in top 10% most aggressive?

Does the potential subject appear to be an adult speaking English, with a Twitter account at least 2 months old?

Is the incivility directed at someone besides a journalist or other political actor?

Is the incivility directed at someone who expressed a different political viewpoint?

Assign to a treatment condition subject to balance constraints

No

EXCLUDE

EXCLUDE

EXCLUDE

EXCLUDE

EXCLUDE

This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded.
few followers had very little effect; several subjects even mocked the bots for having few followers. In the current experiment, I used the same “brand promotion” website to purchase 500 followers for each of my four bots. For reasons I do not understand, the service actually provided each bot with just over 900 followers; the number did not vary significantly among the four.

I created each bot in January 2015, giving the impression that they were long-time users. When creating the accounts, I followed Twitter’s recommendation to follow 40 pre-selected accounts, mostly celebrities and news services. To further increase the perception that the bot was a real person, I tweeted dozens of innocuous observations (eg “I’m thinking of pasta for lunch.....YUM”) and retweeted random (non-political) stories from the accounts the bots followed.

The primary outcome of interest was how subjects responded to being sanctioned, both in terms of their direct response to the sanctioning tweet and in how they changed their behavior after having been sanctioned. I only used bots that appeared to be on the same “side” as subjects to send the sanctioning message; I was concerned that cross-ideological sanctioning might cause subjects to react angrily and send even more uncivil messages. I had no theoretical expectation as to whether right-leaning or left-leaning subjects would respond more to being sanctioned.

However, the within-side variation in treatment effects allows me to test the relative

---

6Interestingly, the price for 500 followers was $1 in Summer 2015, but the same website was charging $10 for the same service in Summer 2016. Other follower-selling sites had similarly increased their prices.

7Bizarrely, the followers I bought sometimes “liked” and even occasionally retweeted these observations, suggesting that at least some of them are real people.
Figure 3: (a) Example Bot–Clinton Condition

(b) Example Bot–Republican Condition
party allegiance of Trump and Clinton supporters. This hypothesis is not based on any long-standing theory in the literature, but on idiosyncratic observations about the 2016 election: Donald Trump’s candidacy has divided the right, and many of his supporters are openly disdainful of the Republican Party.

**Hypothesis 1** The reduction in incivility caused by the Trump condition will be larger than that caused by the Republican condition. There will be no differential effects of the Democrat and Clinton treatments.

The primary variation in the treatments is in the language of the message sent to the subjects. The aim is to convince subjects that their behavior is wrong—or at a minimum, to convince them to change their behavior. One approach, the one I used in a previous experiment with bots in Twitter, is social norm promotion: to cause subjects to update their beliefs about correct normative behavior. I believe that this approach is less appropriate in the current context: as I argue above, incivility is well established as normatively acceptable in arguments on Twitter, and it is unlikely that a single intervention could cause subjects to update their views on this norm.

Instead, I attempted to use moral persuasion. I based my approach on the moral intuitionist model proposed by Haidt (2001), which argues that moral emotion is antecedent to moral reasoning. People make moral judgments based on deep-seated intuitions and then justify those judgments with ad hoc reasoning. As a result, moral appeals should be targeted to these fundamental intuitions, rather than to the putatively logical justifications for specific judgments.

As Haidt argues convincingly in *The Righteous Mind* (Haidt, 2012), a necessary component for moral suasion is convincing your interlocutor that you are sympathetic and understanding. If the two of you share the same fundamental moral intuitions, you can reasonably discuss specific implications of those foundations, but if not, attempts to change their minds are likely to be interpreted as attacks on their worldview and to be met with resistance. To this end, all of my messages begin by identifying my bot and the subject as members of the same party (Democrat/Republican).

Haidt also finds that the morality of liberals and conservatives rests on different foundations. He finds six dimensions of morality that seem to operate in cultures around the world: Care, Fairness, Liberty, Loyalty, Authority, and Sanctity. For an action to fall in the realm of morality, it must either violate or uphold the principles of these moral foundations. He argues that people in non-Western societies are similar to conservatives in the West in that both groups appear to place significant weight on
all six of these moral foundations. Westerners on the left of the political spectrum, however, appear to put far more emphasis on just two: Care and Fairness.

As a result, liberals and conservatives speak past each other on some moral issues. For example, liberals sometimes have difficulty understanding why conservatives are so upset about flag burning. Burning a flag does nothing to cause harm (the primary question underlying the Care foundation), nor is it unfair, so liberals tend not to see it in moral terms. Conservatives, though, feel that it is disloyal and disrespectful to authority, and that flag burning is thus immoral.

To effectively engage in moral suasion, then, you must appeal to the correct moral foundation of your interlocutor. To that end, I designed two different treatments. The first was designed to appeal to the Care foundation, and thus to have some effect on conservatives but a much larger effect on liberals:

@[subject] You shouldn’t use language like that. [Republicans/Democrats] need to remember that our opponents are real people, with real feelings.

The other treatment appealed to the Authority foundation. My expectation was that it should have an effect on conservatives but not on liberals:

@[subject] You shouldn’t use language like that. [Republicans/Democrats] need to behave according to the proper rules of political civility.

In addition to these moral foundations treatments, I included a “placebo” treatment. The goal was to separate out the effect of being tweeted at by a stranger from the specific moral suasion of the main treatment tweets. My intention was to use a message that would serve to remind subjects that their uncivil tweets were public, and my (pre-registered) hypothesis was that this treatment would decrease the subjects’ use of incivility, but that the effect would be smaller than the moral treatments. However, I have noticed that several subjects interpreted the language as a veiled threat, and I’m currently unsure what to expect from subjects who received this message:

@[subject] Remember that everything you post here is public. Everyone can see that you tweeted this.

**Hypothesis 2** The reduction in incivility caused by the Care condition will be larger for liberals than for conservatives. There should be a reduction in incivility caused by the Authority condition for conservatives, but not for liberals. There should be a reduction in incivility cased by the Public condition, but it should be smaller than the other effects.
My subjects were labeled “liberal” and “conservative” above based on which side of the political spectrum they were being uncivil to. I incorporate a technique developed by Barberá (2015) to estimate the ideology of my subjects in a continuous space based on the network of who they follow on Twitter. This technique is well-validated for people who are even moderately interested in politics, which is likely to be the case for my subjects. To the extent that the different moral appeals operate on liberals and conservatives, the effects should be more pronounced on more ideologically extreme subjects.

**Hypothesis 3** The differences in effect sizes for liberals and conservatives specified in Hypothesis 2 should be larger for more ideologically extreme subjects.
References


Gulati, Jeff, and Christine B Williams. 2010. “Communicating with constituents in 140 characters or less: Twitter and the diffusion of technology innovation in the United States Congress.” Available at *SSRN 1628247*.


Phillips, Whitney. 2015. This is why we can’t have nice things: Mapping the relationship between online trolling and mainstream culture. Mit Press.


